Lazy Learning Associative Classification in MapReduce Framework

S.P.Siddique Ibrahim, M. Sivabalakrishnan, S.P. Syed Ibrahim

Abstract: ***he core objective of the work is to propose a distributed environment based lazy learning Associative Classification (AC). Associative Classification is a hybrid version of data mining tasks which integrated both Association Rule Mining (ARM) and Classification technique to construct accurate classifier. Unfortunately, the AC used for learning these classifier are less popular in real time for building application due to its higher computation time complexity and memory constraints in large volume of datasets. Moreover, single processor's CPU resources and memory are limited, which makes the algorithm incompetent to handle such datasets. To overcome such downsides, we proposed a distributed and parallel computing for lazy learning associative classification for accelerating algorithm performance by projecting the testing instances with large training datasets. In this work, we have implemented MapReduce based algorithms which reduce the computation by eliminates the need of constructing generalized classifier. It also well handled rare rules and generated institutive rules. The proposed algorithm may be suitable in area such as network intrusion detection, fraud detection, crowd analysis, rare disease prediction and crime analysis. Our algorithm has been compared with well known existing algorithms in relations of precision and running time. The experiments result has strengthened the proposed algorithm well handle the rare rules in distributed environment and is making better performance even the size of the datasets is huge.***

Keywords: Association Rule Mining, Rare rules, Lazy Learning, Associative Classification.

I. INTRODUCTION

In the modern decade, the development of PC advances and its applications, for example, databases and its reinforcement stockpiling frameworks has empowered capacity of the high volume of information from both people and machines, especially CCTV camera, and sensors. Big data brings attention in both the academic and the industrial world. Efficient and effective data analytics tools are increasingly required to handle such huge data. The task of classifying such these huge data is a most challenging job. The data mining is the process of find out the hidden patterns k and knowledge through enormous datasets to bring intelligent pattern and logical rules that will help the user to sort out the problems with simple steps. The significant undertakings of information mining are Association Rule Mining (ARM) and Classification are real functionalities and has been utilized in an assortment of uses like web investigation, monetary promoting, target showcasing and so onwards.

Revised Version Manuscript Received on 25 November, 2018.

S. P. Siddique Ibrahim, Assistant Professor, Kumaraguru College of Technology, Coimbatore, India.

Dr. M. Sivabalakrishnan, Associate Professor, School of Computing Science and Engineering, VIT University, Chennai, India.

Dr. S.P. Syed Ibrahim, Professor, School of Computing Science and Engineering, VIT University, Chennai, India.

The ARM is named unsupervised realizing where there is no class name is associated with finding regular guidelines used to discover relationships, association, frequent and infrequent mining, pattern correlations from various kind of databases [1]. It is defined in the form of A -> B, which means a rule encompasses A then it will liable to contain B as well. For Instance: In an e-commerce record, an ARM as: Mobile phone, Head phone -> Power bank means that if the customer buys a mobile phone and headphone together then, she/he will likely to buy a power bank also.

The ARM uses to support and certainty measure to control the standard age. For the most part this system appropriate for frequent itemset mining [2]. Grouping is an information mining capacity that relegates classes to an accumulation of information so as to help in increasingly exact expectations and examination. The point of order is to precisely compute the objective class for each case in the information. For Instance, a grouping model could be utilized to recognize Visa candidates as low, medium, or high credit dangers. A grouping undertaking begins with a dataset in which the class marks are known. For instance, an order display that predicts stack application could be produced dependent on watched credit information for some advance candidates over some undefined time frame.

In the preparation stage, a classification technique discovers relationship between the associations between the target and values of the predictors objective and estimations of the indicators. Diverse classification algorithms utilize distinctive strategies for discovering affiliations. These affiliations are abridged in a last model, which would then be able to be connected to an alternate informational index in which the class names are obscure. Some imperative arrangement calculations are choice tree, classifier, insights and Naive-Bayes [22]. They utilize avaricious inquiry systems to discover the subsets of tenets to discover the classifiers.

Big data, the large collection of datasets in the area of social networks, sensor networks, bioinformatics, medicine, gaming, biological, social media, internet and often interdisciplinary fields, are collected in every second of huge data and stored in a common repository. In the year 2020 it has been claimed that Google produces 200PB per day. Knowledge discovery and statistical analysis of big data can be done with data mining tools.

Association also be used with classification process for better classification performance [4,5,6] known as Associative Classification (AC) integrates both ARM and classification [3]. It is rule-based classifier. Here the classifier is developed by ARM worldview, whose subsequent is the class name.

This strategy fills in as pursues: by and large, an AC works in three stages. In the principal stage, it discovers all the arrangement of order affiliation rules (CARs) is produced from the preparation set. Also, rearrange and rank the principles dependent on help and certainty measure. At long last, it builds the classifier and afterward foresee the future variable and compute the calculation exactness. Besides, extraordinary works [7,8,9,10] has been featured that ACs can accomplish high arrangement execution and more exact than customary calculation [11]. It has a straightforward procedure like on the off chance that rules, along these lines enabling the client to effectively decipher and comprehend the calculation [12,13].

Plastino and Merschmann [23] have proposed AC working in the following two ways 1. Lazy Learning and 2. Eager method. In this eager method initially builds classifier and then predict the class label whereas lazy learning methods delay the prediction phase since the fresh model needs to be classified and it will not generate class rules for all instance and it will project the data in the training data sets only on those features based on the test instance. Syed et al., [24] suggested lazy learning in associative classification method which followed information gain novel measure that significantly produces better classification accuracy with less computation cost.

A. MAP REDUCE

Hadoop MapReduce is a programming framework [14] under Apache Hadoop for providing scalability across thousands of cluster systems proposed by Google in 2014. The MapReduce framework process huge datasets and well maintain failures, better utilization of network bandwidth and storage areas. The MapReduce separates the computational stream into two principle stages; namely map and reduce job. Mapreduce is a kind of processing methodology for distributed environment that is based on java platform. It takes the small group of jobs initial datasets as inputs and then process them to produce another set of data in the form {key/ value} pairs. It considers set of the tuples as input key, value pairs and response a set of output same set of pairs. The final results of both the map and reduce jobs could be stored in Apache HDFS.

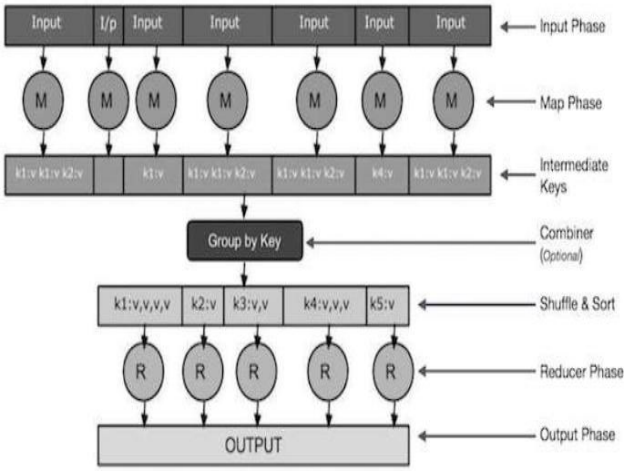

Figure 1.1MapReduce General working model

At first, MapReduce condition executes client program, the structure consequently parcels the datasets into an arrangement of 'n' parts that can be prepared separately by various group machines parallel. Besides, the quantity of guide errands is dictated by the quantity of info parts n, however the quantity of diminish undertaking is characterized by the client. Accordingly, there are n delineate and R diminish errands to be executed without fail. Likewise, the MapReduce system takes duplicates of the client program in various group machines. One of the bunch machine can be chosen as ace hub that calendars and handles assignments inside the groups. The rest of the machines are called laborers hub. The fundamental ace hub allots an undertaking to any sit laborer bunch hub. At that point the specialist hub parses the contribution as {key, value} combines and passes the computational stream to delineate dependent on client characterized program. Frequently the specialists spare the halfway estimations of guide work into neighborhood stockpiling and parcel these into R parts. At last, these specialist returns back outcomes to the ace hub. In the meantime, the reducer hubs read remote halfway information from the nearby plates of guide hub, and the sorts and solidify the significant records from mapping yield. Lately, a few open source instruments and ventures have been created to manages such huge information. Apache start [15], Apache S4[16], Strom [17], Dreme[18], Apache Drill [19], Vowpal wabbit [20], MOA[21] and the organization such Facebook, Twitter, Yahoo!.

The inspiration of the proposed work lazy learning associative classification decreases the critical step in the class rule generation and lessen the calculation time by running the class projection task in distributed environment.

The left of the paper is composed as pursues: Section 2 portrays the Technical fundamentals for the proposed work.

In segment 3 will introduce proposed distributed version of lazy learning technique is clarified. In section 4 provides experiment result computation that gives details about how the proposed algorithm deals both frequent and rare rules.

The final section present with different analysis of the proposed work followed by the conclusion.

II. TECHNICAL PRELIMINARIES

Let T be the set of arrangement of sample dataset with i cases is spoken to by < AT_1, AT_2 …AT_m, C> clolumns, and |R| rows. Where AT_1, AT_2 …,AT_m are characteristics and given C a chance to be a rundown of class Labels. A thing is characterized by the relationship of characteristics and its incentive in frame (ATi, ai) (or) mix of any an incentive among I and m. A typical standard r is characterized as a - > b, where the forerunner of the standard an is a thing and b ensuing is a class mark.

The presence of the rule in the informational collection of the rule r in T is the occasions the rule r has been available in T. The support value s of r is the quantity of cases in T that matches r's forerunner, with class label C. A rule r needs to pass the min support threshold(min_supp) if for r,

the supcount (r)/|T| ≥ minsupp, a rule r needs to pass the minconf if supcount (r)/Appr (r) ≥ minconf

Table 1 Training Data

ID	CT_1	CT_2	Class
100	T_1	T_3	C_1
200	T_1	T_4	C_2
300	T_2	T_5	C_2
400	T_1	T_5	C_1
500	T_2	T_3	C_2
600	T_2	T_4	C_2

III. PROPOSED ALGORITHM

In Existing methods, the training phase of an associative classifier is a memory- exhaustive process, often it executed out of core. huge arrangement of itemsets amid standard extraction stage and not yet pruned. This proposed model can't use the benefits of our reference design, an in-memory bunch figuring system like Hadoop. The MapReduce architecture adopted for proposed Lazy Associative Classification shown in Fig. 1. According to architecture the data collected from different sources are split into n and stored in Hadoop distributed file system (HDFS). The HDFS runs MapReduce model controlled by Name Node where the Associative program for processing the data will be submitted. The Task-tracker function of the name node will schedule the map task by initiating by map nodes based on the data to be processed. The data at initiated Map nodes will be processed locally in parallel mode. Once the entire map job has been done the local intermediate results will be shuffle and locally saved and then forward to Reducer stage where solidified preparing will be consented to create the last

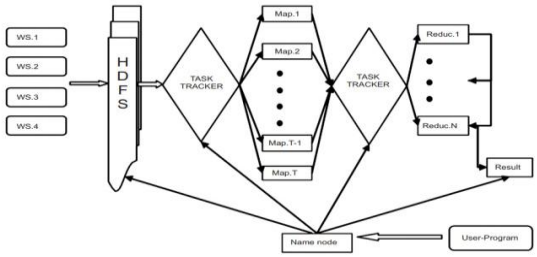

Figure 3.1MapReduce Architecture

outcome. The quantity of reducer capacities to be made will be chosen by the client program rationale for proficient changing of moderate outcomes into overall outcomes.

The steps of general Lazy Learning algorithms are furnished below:

Step 1: Parts the preparation informatoin into n splits

Step 2: Assign a testing sample to name node

Step 3: Project the instance in the training data set stored in HDFS

Step 4: Class rule mapping process with training dataset during mapper function.

Step 5: Predict the new class instance in the reduce phase

Step 6: Calculate the accuracy

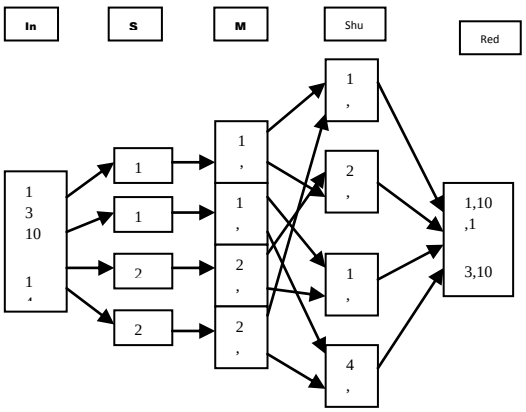

Figure 3.2MapReduce proposed Architecture

IV. EVALUATION AND RESULTS

The classifier predicts the class of each experiment: in the event that it is right, it is considered a hit; if not, it is a misstep. The mistake rate is only the level of blunders made over an entire arrangement of experiments, and it gauges the opening execution of the classifier. The 10-fold cross-approval for the most part utilized in classification algorithm. First the information is separated into 10 sections arbitrarily in which the class is spoken to in roughly indistinguishable extents from in the full dataset. Each part is seized out thus and the learning plan prepared on the staying nine-tenths; at that point the mistake rate is determined on the holdout strategy. Therefore the learning method is executed a sum of multiple times on various preparing sets.

V. EXPERIMENT RESULTS

The proposed framework was tried utilizing a few UCI [25] well known benchmark datasets, specific experimental tests have been devised the execution of the proposed method based on the accompanying angles; i) Computation time ii) Classification precision iii) Scalability examination. As appeared in the below Table 2, we utilized 7 well understood datasets. All the datasets has been executed in Hadoop stage. In the dataset, Holdout approach [26] was utilized where 90% of the datasets are utilized as preparing information and 10% of the datasets are utilized as testing information.

TABLE 2 DATASET DESCRIPTION

Dataset	No. of Attributes	Instances	Classes
COV	54	581012	2
HIGGS	28	11000000	2
KDD 99_2	41	4856151	2
KDD 99_5	41	4898431	5
POK	10	1025010	10
SUSY	18	5000000	10

We utilized the following rule matching procedure for classifier construction:On the off chance that conf (Ri) > conf (Rj) Or on the other hand conf(Ri) = conf(Rj) and supp(Ri) > supp(Rj) Of conf(Ri)=conf(Rj) and supp(Ri)=supp(Rj) and length(Ri) < lenth (Rj).At that point the calculation picks the standard with the most astounding position and sent it to foresee the class name of the new object.

All the experiments were carried out on a system with Ubuntu14.04 Operating System, Intel dual core CPU with 3.4GHz clock speed and 8 GB of primary memory.

A. Accuracy Calculation

Accuracy measure is the limit of a classifier to accurately arrange unlabeled information. It is the count proportion of the quantity of effectively classifier information over the aggregate number of exchanges in the test dataset.

$$\text{Accuracy} = \frac{\text{Number of known values in the test data}}{\text{Total Number of instances}}$$

Table 3 illustrations lazy MapReduce associative classification has greater accuracy than existing classification algorithms.

TABLE 3 ACCURACY COMPUTATION

Dataset	Proposed Algorithm	MRAC	Decision Tree
COV	79.98	78.33	75.83
HIGGS	69.45	67.30	66.24
KDD 99_2	99.98	99.56	99.97
KDD 99_5	99.6	99.32	99.78
POK	95.67	94.65	56.16
SUS	80.12	78.34	76.11

According to the Table 4 RLLAC improves the running performance of classification when compared with Existing algorithms.

TABLE 4 COMPUTATION TIME

Dataset	Time taken by Proposed algorithm	MRAC
COV	550	600
HIGGS	7634	7729
KDD 99_2	770	780
KDD 99_5	1532	1604
POK	249	302
SUS	873	932

In figure 5.1 a chart is drawn contrasting the count of rules produced through MapReduce LLAC and other conventional algorithms. The x pivot demonstrates the different sorts of algorithms and the y pivot demonstrates the total number of rules produced. What's more, unmistakably the proposed MapReduce based LLAC calculation produces more arrangement of helpful rules than customary algorithms.

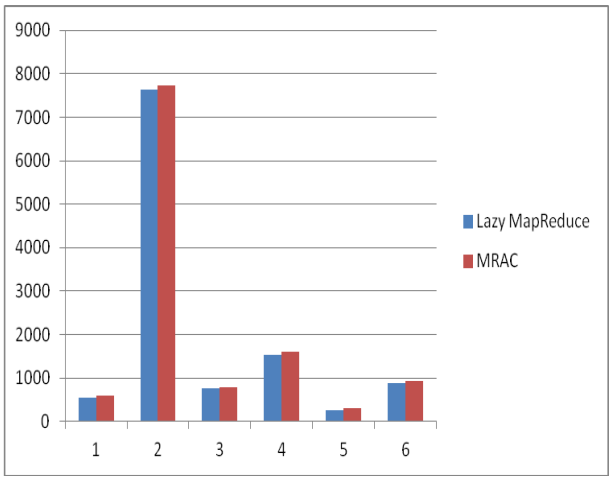

FIG. 5.1. COMPARISONS OF DIFFERENT RULE MINING ALGORITHMS

VI. CONCLUSION

In this paper, we proposed a MapReduce lazy learning associative classification. The distributed environment has been taken for this experiments since the Hadoop is highly scalable and it ability to store huge data as well as distribute large datasets across various cluster machines. These cluster machines are inexpensive and able to operate in parallel. The main idea for lazy associative classification is to assemble computationally productive classifier. The proposed technique predicts the class for each example dependent on assessment of its subsets of quality qualities. Experiments results performed on six datasets shows that proposed lazy associative classification using MapReduce framework well handle the big datasets and produced better accuracy and computation complexity than existing algorithms.

As future work, we intend to tests the proposed work in Apache Spark platform by large cluster of machines.

REFERENCES

1. Agrawal R. and Srikant R. “Fast algorithms for mining association rule” Proceedings of the twentieth International conference on very large databases. 1994, pp. 487-499.
2. B. Liu and Y. Ma, “Mining association rules with multiple minimum supports,” in Proc. fifth ACM International Conf. Knowledge discovery data mining, 1999, pp. 337–341.
3. Apache Drill, http://drill.apache.org May 2015).
4. Apache Spark, https://spark.apache.org (Accessed: May 2015).
5. Apache Storm, https://storm.apache.org (Accessed: May 2015).
6. M. I. A. Ajlouni, W. Hadi, J. Alwedyan, Detecting phishing websites using associative classification, European Journal of Business and Management 5 (15) 2013, pp. 36-40.
7. N.Abdelhamid A, Ayesh .F “A Multi class Associative Clasfication algorithm” Information Knowledge Management, 2012.
8. E. Baralis, P. Garza, “ A Lazy approach to pruning classification rules” Proceedings of IEEE Internaltional conference on DataMining, 2012, pp. 35-42.
9. B. Liu, W.Hsu, et.al., “Integrating classification and association rule mining”, In proc. of the Fourth International Conference on Knowledge Discovery and Data Mining, 1998, pp. 80-86.

10. Baralis. E, Chiusano. S, A Lazy Approach to Associative Classification, IEEE Transactions on Knowledge and Data Engineering, v.20 n.2, p.156-171, February, 2008.
11. Quinlan. J, "Induction of decision trees". Machine Learning, pp, 81–106, 1986.
12. Han. J, and Pei. J, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules," Proc. of IEEE International Conference on Data Mining (ICDM '01), Nov. 2001.
13. [Guoqing Chen, et.al., "A new approach to classification based on association rule mining", Science Direct, Decision Support Systems 42 ,2006, pp. 674– 689.
14. J. Dean, S. et.al., "A MapReduce: A flexible data processing tool", Comm. ACM 53 (1), 2010, pp.72–77.
15. Apache Spark: October 2015. (Https:// spark.apache.org)
16. L. Neumeyar, B.Robbings, " Distributed stream computing platform" Proc. of IEEE Int. conference on data mining workshop, 2010, pp. 170-177.
17. [Apache Storm: October 2015. (https:// storm.apache.org)
18. S. Melnik, A.Gubarev, " Interactive analysis of web-scale datasets", Proc. VLDBndow, 2010, pp. 330-339.
19. Apache Drill: October 2015. (http:// drill.apache.org)
20. Vowplawabbit.:October 2015(http://hunch.net/~vw/).
21. Bifet,G.Holmes, "Massive online analysis", 2010, pp. 1601-1604.
22. Wu, X., et.al. 2007.Top 10 algorithms in data mining. Knowledge Information System, 2008, pp. 1-37.
23. Cendrowska J. "An Algorithm for inducing modular rules" Internal. Jour. of Man-Machine Studies. Vol.27, 1987, pp. 349-370.
24. Syed Ibrahim. S.P, Nataraj R.V., "LLAC: Lazy Learning in Associative Classification" in the Springer Lecture Series (CCIS) Part I, 2011, PP. 631 – 638.
25. UCI Machine Learning Repository: Data Sets [Online] (2010). Available: http://archive.ics.uci.edu/ml/datasets.html.